What Clarity Costs: Measuring Legible Robot Motion Against Path Cost and Constraint Satisfaction

Munawar Kazmi
Independent
United Kingdom
munawarsaeedkazmi1@gmail.com

Abstract

A robot moving towards one of several possible goals is unreadable for the first seconds of its motion, which is when a person nearby hesitates or steps into its path. Legible motion deviates early to resolve that ambiguity, and the deviation costs path length. What it can also cost is the constraint the robot was supposed to respect. We present a small judge-free benchmark that measures the three together: legibility in the form of Dragan et al., path cost against the exact optimum, and satisfaction of stated keep-out constraints. Eight worlds each carry machine-checked facts about their own geometry, optimal cost-to-go is computed exactly over a visibility graph, and no human rater or model judge appears anywhere in the scoring loop. Asked to produce legible motion under a stated path budget, three language models returned 120 trajectories, and they differ exactly where it matters. Two of them called all 80 of their trajectories legible, including the 25 that were not physically possible, and broke the stated budget 24 times between them; in the world built so that the clearest signal is the one the robot is not allowed to send, both entered the keep-out zone on every one of ten samples. The third stayed inside the budget and outside the zone on all 40, and beat the shortest path in that same world on five of five without entering it. Given four different path budgets the first two moved the cost they actually paid by 0.012 and by nothing at all; given two, the third moved with the budget in seven of eight worlds. Its four refusals turn out to be about the stated budget rather than the motion: the same trajectory in the same world is called not legible under a tight budget and legible under a loose one.

CCS Concepts

• **Computer systems organization** → **Robotics**; • **Computing methodologies** → *Motion path planning*; • **Human-centered computing** → *Human computer interaction (HCI).*

Keywords

legibility, human-robot interaction, motion planning, benchmarking, language models

1 Introduction

Legibility is not ours. Dragan et al. [3] formalised it: a legible trajectory is one that departs from the efficient path early enough for an observer to infer the goal sooner than efficiency alone would allow. What is missing from that literature is a price. Deviating for clarity moves the robot somewhere it would not otherwise go, and in a real room that somewhere is occupied.

The literature notices this and leaves it qualitative: what a legible trajectory costs in satisfaction of a stated constraint is not an axis of any reported result.

This report contributes an instrument rather than a planner: a reproducible way to ask what a given trajectory pays, in path and in constraint satisfaction, for the clarity it buys. We use it to ask whether language models produce legible motion when told to.

2 Related work

Legibility in the sense used here is Dragan et al.'s: an observer with a Boltzmann-rational model of the robot infers the goal from motion, and a legible trajectory is one that raises that inference early [3]. Both founding papers bound what the robot may spend on it. The HRI paper subtracts a cost regulariser from the objective and the RSS paper constrains the trajectory to a trust region, in both cases because a robot can otherwise make a trajectory ever more legible at ever greater cost [4]. The ceiling we sweep is the hard variant of that bound, expressed as a ratio against the optimal path rather than as an absolute in a cost functional.

What legibility does to the space around the robot is stated in that literature before it is measured. The RSS paper reports that legibility moves the trajectory much closer to an obstacle in order to disambiguate between two goals, and leaves it there: obstacles enter as a soft penalty inside the objective, clearance is never reported as a number, and constraint satisfaction is not an axis of any result [4]. Legibot likewise carries obstacle distance and rate of approach as terms in a task cost, reports legibility scores from a user study, and gives no quantitative comparison of path cost against the baseline [1]. Closest to this report is legible navigation among moving agents, which does put a safety quantity beside legibility in one table, reporting minimum separation and minimum time to collision [2]. What differs here is which quantity and in what form: satisfaction of a stated static constraint rather than proximity to moving agents, traced as a curve against a path budget rather than compared between policies.

On evaluation, the guidelines for social robot navigation call for repeatable benchmarking criteria, endorse algorithmic metrics as cheap to compute and reproducible, name legibility as a principle, and point at Dragan for its definition rather than stating a computed metric of their own [5]. They also set the condition this instrument does not meet, that objective metrics be validated empirically, which the limitations take up. The alternative is to evaluate legibility frameworks against human observers on framework-independent trajectories [7]. That is the harder experiment and it measures something this one does not: an exactly computed observer is reproducible rather than right.

Language models have been asked for expressive robot behaviour before. GenEM prompts GPT-4 to reason about social context and emit control code for a mobile robot, producing behaviours such as nodding or excusing itself through speech, body movement and a light strip [6]. The pairing is the same and the quantity is not. Legibility in Dragan's sense does not appear in that work, no observer infers a destination anywhere in it, and the behaviours are judged by participants watching video clips on seven-point scales against a professional animator's versions. What is asked for here is narrower and checkable: a trajectory under a stated path budget, in a world whose optimal cost-to-go is computed exactly.

3 The instrument

Everything is 2D and kinematic. A trajectory is a sequence of positions traversed at constant speed. There is no physics engine, which is a scope decision: physics would improve the pictures and change none of the numbers.

The observer is Boltzmann-rational in exactly the form of [3], equation 8, and our legibility metric is their equation 9 including the weighting $f(t) = T - t$. Both were compared line by line against the implementation. Our only addition is an inverse temperature, which is recorded in the observer's name in every result.

Two observers are implemented and every number here uses the first. One takes cost-to-go to be the obstacle-aware geodesic; the other takes it to be the straight line, modelling someone who can see the robot and knows the candidate goals but not what stands between them. That is a measurement decision rather than an ablation, and it changes the measurement. Along the cheapest route in the world with a wall to round, the informed observer holds at the prior of exactly 0.5000 and learns nothing, while the naive one falls to 0.3164, because from a viewpoint that cannot see the wall the robot is walking away from the goal it is going to. Legibility on that one path is 0.5457 read the first way and 0.4370 read the second.

That posterior rests on the optimal cost-to-go, so we compute it exactly rather than approximately. Obstacles are convex polygons, the shortest obstacle-avoiding path is a polyline through their vertices, and the optimum comes from search over the visibility graph: no grid, no resolution to defend. The geometric predicate underneath is guarded, decided in floating point wherever an error bound proves the sign cannot have changed and in exact rational arithmetic otherwise. On 20,000 near-collinear cases an unguarded floating point determinant returns the wrong sign on more than a tenth of them; the guarded predicate matches rational arithmetic on all of them, and the test suite asserts both halves so the corpus cannot quietly become easy.

Keep-out zones do not block motion. A trajectory may cross one and is scored for having done so. If they blocked motion there would be no trade to measure.

Legibility optimisation is bounded by a ceiling on the cost ratio. That bound is not our idea: [3] adds a cost regulariser and [4] a hard trust region, both because a robot can otherwise make a trajectory ever more legible at ever greater cost. Sweeping the ceiling turns a single trajectory into a frontier.

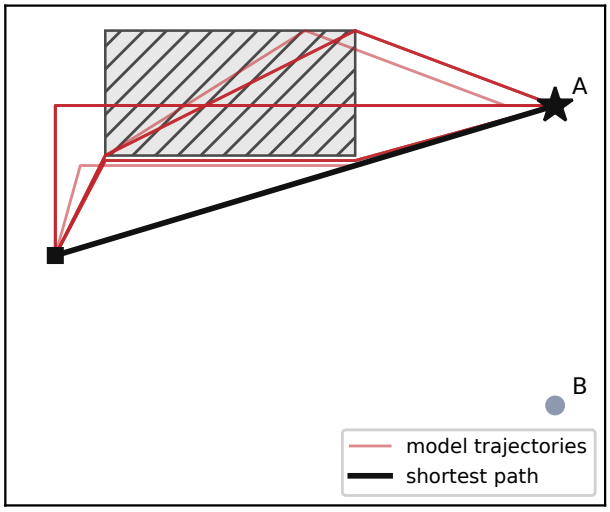

Figure 1: The world built so that the clearest signal is the one the robot may not send. The cheapest route to the true goal A (black) stays clear of the keep-out zone (hatched), so safety costs nothing until a planner tries to be legible. All fifteen trajectories the three models returned (red, five samples each) are more legible than the shortest path. Ten of them cross the zone and five do not, and the five that stay out are one model's, on five of five samples. Drawn from the committed records, not by hand.

4 What language models do

The prompt states the room, the goals, which goal the robot is going to, the obstacles, the keep-out zones, and a path budget in absolute units. It asks for a trajectory that makes the destination clear as early as possible, and asks the model to judge its own answer. Temperature 0.7, five samples per world, eight worlds.

What the three models share is thin, and what separates them is not.

They agree in one world. Where the true goal lies between the other two, all fifteen decodes were physically possible and none beat the shortest path, which is what [3] predicts when exaggeration towards a middle goal points at a different goal.

Elsewhere the differences are the result. In the world with a wall to round, one model returned a physically impossible trajectory on all five samples, one on one of five, and one on none. In the world built so that the clearest signal and the keep-out zone conflict (Figure 1), all three beat the shortest path on five of five. Two of them entered the zone on five of five; the third entered on none, threading a route under the boundary at a cost ratio between 1.0819 and 1.1057. The constraint was satisfiable with the clarity still available, and the first two give no sign of having looked for that.

The claim behaves the same way, and a second budget shows what it is worth. Two models called every one of their 80 trajectories legible, including 25 that were not physically possible. The third declined four times at a stated budget of 1.25, three in the middle-goal world and one in the world with the wall, which reads like

Over 40 decodes each	Gemini 3.6 Flash	Llama 3.3 70B	Qwen 2.5 7B
replies that parsed	40	40	40
called legible by the model	36	40	40
physically possible	40	29	26
clearer than the shortest path	30	20	10
over the stated path budget	0	15	9
entered a keep-out zone	0	7	7

Table 1: Counts at a stated path budget of 25 per cent. Of 120 decodes, 116 claimed to be legible, including the 25 that were not physically possible. All four refusals came from one model, and none of them survives raising the stated budget to 2.00.

	Gemini 3.6 Flash		Llama 3.3 70B		Qwen 2.5 7B	
Stated budget	median cost	over	median cost	over	median cost	over
1.10	-	-	1.2817	30	1.1663	17
1.25	1.0819	0	1.2817	15	1.1588	9
1.50	-	-	1.2817	10	1.1712	3
2.00	1.1539	0	1.2817	0	1.1709	0

Table 2: Forty decodes at each stated budget, per model. For the two models swept across all four, the budget nearly doubles and the cost actually paid does not follow it. The third was run at 1.25 and 2.00, where it does follow. A dash is a cell that was not run rather than a cell with nothing in it.

a model that knows when deviating cannot help. It is not that. In the middle-goal world it returns the identical trajectory under both budgets, the straight line through $(1, 5)$ and $(11, 5)$, scoring exactly the baseline. That trajectory is called not legible on three of five samples at 1.25 and legible on five of five at 2.0. Same world, same motion; only the number in the prompt changed.

Table 2 asks whether the budget violations mean the budget was too tight or that the budget was not attended to. For the two models swept across all four budgets it is the second. One model's median cost moves by 0.012 across a budget that nearly doubles and the other's does not move at all, so the falling violation counts are the line moving, not the models complying. Where the optimiser sits exactly on the budget in every row, because for a planner it is a constraint, for those two it is text.

The third model separates the two readings that compliance at a single budget cannot. Given 2.0 rather than 1.25 it spends more, in seven of the eight worlds, and its pooled median rises from 1.0819 to 1.1539. Where the extra room buys something it takes much more of it: in the keep-out world the median cost goes from 1.0819 to 1.4139 and legibility with it, 0.7710 to 0.8496, without a single keep-out entry at either budget. In the world with a wall to round it crosses from below the shortest path's legibility to well above it, 0.5339 to 0.7575 against a baseline of 0.5457. In the world where the true goal lies between the others it returns the same straight line under both budgets and spends nothing extra. It never exceeded either stated budget.

5 Limitations

Validity is the limitation that matters, so it goes first. The metric here is not a new claim about people. It is equation 9 of [3] computed exactly rather than approximately, so whatever validity it has is inherited from that paper's user studies, and it is bounded where those authors bounded it, inside the trust region. That is why nothing outside a stated cost ceiling is reported here. What this instrument adds is reproducibility, not correspondence. Judge-free is not judge-proof.

Two statements have to be kept apart, and only the first is supported by anything here: that a trajectory scores higher under the stated observer model, and that a person would read it sooner. Nobody has watched these trajectories. The guidelines for social navigation put the same order on it, recommending algorithmic metrics first and asking that objective metrics later be validated empirically [5]; this report is at the first step and not the second. What would settle it is a study in which people see these trajectories and say which goal they believe, set against the posterior that scored them. Every trajectory is committed, one JSON object per line, so that study needs participants rather than a reimplementation. Comparing legibility frameworks against human observers remains the harder and more informative experiment [7].

Three narrower limits. With more than one goal the legibility score cannot reach 1, so a value of 0.93 is not 93 per cent of anything. Our observer has no option for a goal outside the declared set, though real observers form exactly that belief when motion becomes strange enough. And time to confidence, the one metric here with a free parameter, is not stable: across the admissible band the count of trajectories reaching confidence sooner than the baseline runs from 55 of 95 down to 31 of 95, and individual verdicts move with it. It is reported nowhere above and no claim here rests on it.

Three models at one temperature over eight worlds is a pilot, and one was swept at two budgets rather than four. Every number here is recomputed by a committed tool from committed records.

6 Availability

The eight worlds and the facts each one carries, the prompt template, every record file behind the tables above, and the tools that recompute every table and the figure from those records, are public.

The link is withheld here for anonymous review and will be given at camera-ready. Nothing reported above was produced by a step that is not in that repository.

The instrument is not specific to language models. Anything that emits a trajectory can be scored by it, and the cost of adding a planner or a model is one record file. The worlds and the facts they carry are the fixed part; what gets measured against them is not.

References

[1] Javad Amirian, Mouad Abrini, and Mohamed Chetouani. 2024. Legibot: Generating Legible Motions for Service Robots Using Cost-Based Local Planners. *arXiv preprint arXiv:2404.05100* (2024).

[2] Jean-Luc Bastarache, Christopher Nielsen, and Stephen L. Smith. 2023. On Legible and Predictable Robot Navigation in Multi-Agent Environments. In *IEEE International Conference on Robotics and Automation (ICRA).*

[3] Anca D. Dragan, Kenton C. T. Lee, and Siddhartha S. Srinivasa. 2013. Legibility and Predictability of Robot Motion. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI).* 301–308.

[4] Anca D. Dragan and Siddhartha S. Srinivasa. 2013. Generating Legible Motion. In *Robotics: Science and Systems (RSS).*

[5] Anthony Francis, Claudia Pérez-D'Arpino, Chengshu Li, Fei Xia, Alexandre Alahi, et al. 2025. Principles and Guidelines for Evaluating Social Robot Navigation Algorithms. *ACM Transactions on Human-Robot Interaction* (2025). Also available as arXiv preprint arXiv:2306.16740.

[6] Karthik Mahadevan, Jonathan Chien, Noah Brown, Zhuo Xu, Carolina Parada, Fei Xia, Andy Zeng, Leila Takayama, and Dorsa Sadigh. 2024. Generative Expressive Robot Behaviors using Large Language Models. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI).*

[7] Sebastian Wallkötter, Mohamed Chetouani, and Ginevra Castellano. 2022. A New Approach to Evaluating Legibility: Comparing Legibility Frameworks Using Framework-Independent Robot Motion Trajectories. *arXiv preprint arXiv:2201.05765* (2022).